BINAURAL DARK-VELVET-NOISE REVERBERATOR

Jon Fagerström[1], Nils Meyer-Kahlen[1], Sebastian J. Schlecht[1,2] and Vesa Välimäki[1]

[1]Acoustics Lab, Department of of Information and Communications Engineering
[2]Media Lab, Department of Art and Media
Aalto University, Espoo, Finland
jon.fagerstrom@aalto.fi

ABSTRACT

Binaural late-reverberation modeling necessitates the synthesis of frequency-dependent inter-aural coherence, a crucial aspect of spatial auditory perception. Prior studies have explored methodologies such as filtering and cross-mixing two incoherent late reverberation impulse responses to emulate the coherence observed in measured binaural late reverberation. In this study, we introduce two variants of the binaural dark-velvet-noise reverberator. The first one uses cross-mixing of two incoherent dark-velvet-noise sequences that can be generated efficiently. The second variant is a novel time-domain jitter-based approach. The methods' accuracies are assessed through objective and subjective evaluations, revealing that both methods yield comparable performance and clear improvements over using incoherent sequences. Moreover, the advantages of the jitter-based approach over cross-mixing are highlighted by introducing a parametric width control, based on the jitter-distribution width, into the binaural dark velvet noise reverberator. The jitter-based approach can also introduce time-dependent coherence modifications without additional computational cost.

1. INTRODUCTION

Sound for virtual and augmented reality requires rendering binaural reverberation. Binaural reverberation is characterized by a specific, frequency-dependent interaural coherence (IC), which depends on the soundfield and the physiology of the human head. Early on, artificial reverberation (AR) algorithms have been proposed that aim at matching the broadband IC by matching the maximum of the interaural cross-correlation (IACC) to a measured binaural room impulse response (BRIR) [1]. Later, it was shown that matching the broadband IC is not accurate enough; instead, the frequency-dependent IC should be matched to synthesize binaural reverberation [2].

Menzer and Faller proposed a modified feedback-delay-network (FDN) reverberator with frequency-dependent IC matching [3, 4]. Their method requires synthesizing two uncorrelated outputs of an FDN, which are then filtered and cross-mixed with specially designed coherence-matching filters. An alternative method is to generate several uncorrelated outputs of the reverberator and to convolve each of them with a head-related transfer function (HRTF) belonging to a different direction [5, 6]. Kirsch et al. have found that at least 12 directions are required if the reverberation is isotropic, i.e., it does not depend on direction [7]. However, in the isotropic case, the cross-mixing approach is more efficient.

Karjalainen and Järveläinen proposed Velvet noise, due to its sparsity and perceptual smoothness, as an effective and efficient model of the late reverberation tail [8]. Various feedback-based [8, 9, 10] and feedforward-based [11, 12] velvet-noise-reverberators have been proposed for generating the frequency-dependent decay found in natural late reverberation. The latest in the feedforward-based methods is the dark velvet noise (DVN) [13], which was later generalized as the extended DVN (EDVN). Compared to FDNs and other recursive algorithms, EDVN allows non-exponential decay and simpler matching of the frequency-dependent decay and overall coloration [14].

This paper proposes binaural dark velvet noise (BDVN) as a two-channel version of EDVN [14]. BDVN generates binaural reverberation with a given IC using two variants. The first one employs cross-mixing as in [3, 2]. It is demonstrated how the two required incoherent sequences can be generated efficiently through small modifications of the existing EDVN structure. The second approach jitters one sequence relative to the other, utilizing the fundamental relationship between IC and jitter distribution. As far as we know, this relationship has not been previously employed for synthesizing binaural reverberation.

The rest of this paper is organized as follows. Sec. 2 summarizes the relevant background on IC and the EDVN algorithm. Sec. 3 describes the novel BDVN algorithm and its two variants. Sec. 4 presents objective and perceptual evaluations. Sec. 5 concludes the work.

2. BACKGROUND

2.1. Binaural Reverberation

Many artificial reverb algorithms, including the EDVN reverberator, produce single-channel output. However, in real-world listening, we experience binaural reverberation, with differences in the signals received by each ear. The IC is defined as

$$\Phi_{\mathrm{LR}}(\omega) = \frac{|S_{\mathrm{LR}}(\omega)|^2}{S_{\mathrm{L}}(\omega) S_{\mathrm{R}}(\omega)}, \tag{1}$$

where ω is the frequency in radians, $S_{\mathrm{L}}(\omega)$, $S_{\mathrm{R}}(\omega)$ are the power spectral densities (PSDs) of the two ear channels, and $|S_{\mathrm{LR}}(\omega)|$ is the absolute value of the cross power spectral density (CPSD).

In practice, IC is estimated using time averages of the short-time Fourier-transform (STFT) coefficients

$$\hat{\Phi}_{\mathrm{LR}}(\omega) = \frac{\sum_{\nu=\nu_0}^{\infty} |H_{\mathrm{L}}(\nu,\omega) H_{\mathrm{R}}^{*}(\nu,\omega)|^2}{\sum_{\nu=\nu_0}^{\infty} |H_{\mathrm{L}}(m,\omega)|^2 \sum_{\nu=\nu_0}^{\infty} |H_{\mathrm{R}}(\nu,\omega)|^2}, \tag{2}$$

where $H_{\mathrm{L}}(\nu,\omega), H_{\mathrm{R}}(\nu,\omega)$ are the STFT coefficients of the left and the right channel of the BRIR, respectively, and .* denotes complex conjugation. The sum starts at the time index ν_0. In this study, we focus on modeling only the late part of the response, starting 50 ms after the direct sound arrival—a conservative estimate of the mixing time for most rooms[15]—to estimate IC for the remainder of the work.

If the responses are identical, the IC equals 1 across all frequencies; if perfectly incoherent, the IC is 0. In an ideal diffuse field, sound from different directions is incoherent and has equal intensity. However, the IC measured at a binaural receiver is not zero due to time and level differences caused by the finite size head acting as a scattering object, increasing coherence at low frequencies.

2.2. Cross Mixing

A common method to generate artificial binaural reverberation with a specified IC involves cross-mixing two completely incoherent channel responses, h_1 and h_2 [3, 2, 4]. The mixing filters are

$$U_1(\omega) = \sqrt{\frac{1+\sqrt{\Phi_{\mathrm{LR}}(\omega)}}{2}},\ U_2(\omega) = \sqrt{\frac{1-\sqrt{\Phi_{\mathrm{LR}}(\omega)}}{2}}. \quad (3)$$

They are applied to the incoherent sequences to create the binaural responses

$$H_{\mathrm{L}}(\omega) = U_1(\omega)H_1(\omega) + U_2(\omega)H_2(\omega) \quad (4)$$

$$H_{\mathrm{R}}(\omega) = U_1(\omega)H_1(\omega) - U_2(\omega)H_2(\omega) \quad (5)$$

where $H_1(\omega)$ and $H_2(\omega)$ are the discrete-time-Fourier transforms (DTFTs) of two incoherent sequences $h_1(t)$ and $h_2(t)$, and $H_{\mathrm{L}}(\omega)$ and $H_{\mathrm{R}(\omega)}$ are the DTFTs of the synthesized binaural responses of the left and right ear, respectively.

Menzer and Faller applied the cross-mixing approach for mixing two incoherent Gaussian noise sequences [2] as well as two FDN outputs [3, 4]. Below, we demonstrate the effectiveness of cross-mixing for creating BDVN from EDVN, leveraging the ease of generating two incoherent EDVN sequences.

2.3. Extended Dark Velvet Noise

Original Velvet Noise (OVN), the basis of EDVN, is a sparse pseudo-random sequence composed of a jittered unit impulse train with uniformly distributed signs [8]. As such, OVN has a flat PSD [16]. The placement of each unit impulse is constrained via the grid size T_{d} defined as

$$T_{\mathrm{d}} = \frac{f_{\mathrm{s}}}{\rho}, \quad (6)$$

where f_{s} is the sample rate in Hz and ρ is the pulse density in pulses per second. The sample rate $f_{\mathrm{s}} = 48$ kHz was used throughout this work. A single unit impulse is placed on a uniformly distributed random location within each grid segment. The impulse locations are computed with [8]

$$k(m) = \lfloor mT_{\mathrm{d}} + r_1(m)(T_{\mathrm{d}} - 1) \rceil, \quad (7)$$

where m is the pulse index, $\lfloor\cdot\rceil$ is the rounding operator and $r_1(m)$ is a uniform random number between 0 and 1. The sign of each pulse is given by

$$s(m) = 2\lfloor r_2(m) \rceil - 1, \quad (8)$$

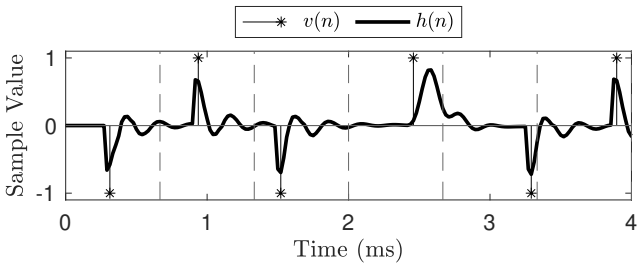


Figure 1: *Beginning of an EDVN IR (line) and the underlying OVN sequence (stems) with a pulse density* $\rho = 1500$ *pulses/s, using sample rate* $f_s = 48$ *kHz.*

where $r_2(m)$ is a uniform random number between 0 and 1. The stems in Fig.1 correspond to the pulses of an OVN sequence.

EDVN is an extension of the OVN to achieve an arbitrary PSD [14]. In the EDVN, each unit impulse of the OVN is replaced with an arbitrary filter IR from a set of $Q \ll M$ predefined dictionary filters F_q, (cf. Fig 2a). Additionally, the pulse signs $s(m)$ and pulse gains $g(m)$, which are used to control the broadband decay, are combined into a single variable

$$g_s(m) = s(m)g(m). \quad (9)$$

Fig. 1 shows the beginning of an EDVN IR (line). Although, the resulting IR is no longer sparse the desired sparse convolution property is retained within the delay line (purple) in Fig 2a.

In Fig. 2a, the qth EDVN sub-sequence containing the pulses routed to the qth dictionary filter is given by

$$v_q(n) = \begin{cases} g_s(m) & \text{for } n = k(m) \wedge \phi(m) = q, \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where $q \in \{1, 2, ..., Q\}$, is the pulse filter index, n is the sample index, $g(m)$ is a gain term for each pulse, and $\phi(m) \in \{1, 2, ..., Q\}$ a list of filter indices. The transfer function of the EDVN is given as

$$H(z) = \sum_{q=1}^{Q} V_q(z)F_q(z), \quad (11)$$

where $V_q(z)$ is the transfer function of the sub-sequence routed to the qth pulse filter with the transfer function $F_q(z)$. An example of a generated sequence $h(n)$ is shown in Fig.1.

The pulse filters are selected based on filter probability, denoted by the vector representing probabilities for a single pulse

$$\boldsymbol{p} = \left[p_1, p_2, \ldots, p_Q\right]^{\mathrm{T}} \geqslant 0, \text{ with } \sum_{q=1}^{Q} p_q = 1, \quad (12)$$

where $[\cdot]^{\mathrm{T}}$ is the transpose operation. The list of filter indices for each pulse is then determined based on the pulse-filter probabilities $\boldsymbol{p}(m)$ with the following greedy selection:

$$\phi(m) = \arg\max_q \{(\tau_q(m) + \epsilon r_q(m))p_q(m)\}, \quad (13)$$

where ϵ is a free parameter for the level of randomization, $r_q(m)$ is a uniform random number between 0 and 1, and τ_q is the sample index (i.e., time) when the qth dictionary filter was last selected.

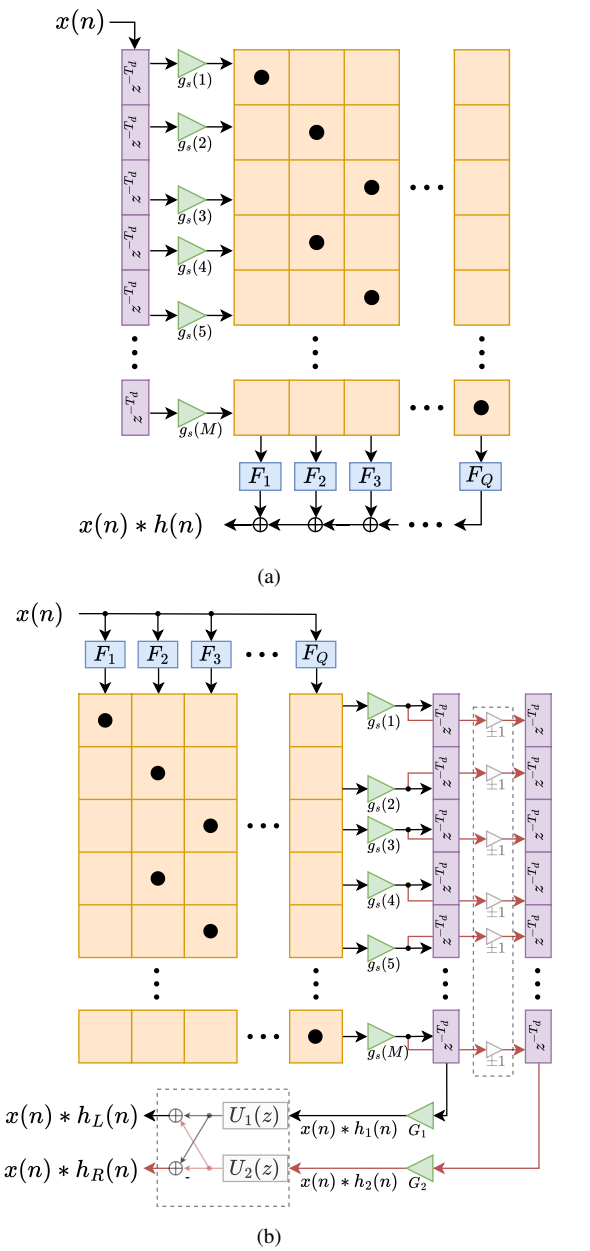


Figure 2: *Structures of (a) the single-channel EDVN convolution and (b) the proposed BDVN convolution with Q dictionary filters and M pulse gains. The translucent blocks (dotted line) are only needed for the cross-mixing version and are omitted when using the jitter version. The ± 1 gains represent the random sign flips.*

The value τ_q is updated sequentially based on the previous pulse filter selection with

$$\tau_q(m+1) = \begin{cases} 0 & \text{for } \phi(m) = q, \\ \tau_q(m) + 1 & \text{otherwise.} \end{cases} \tag{14}$$

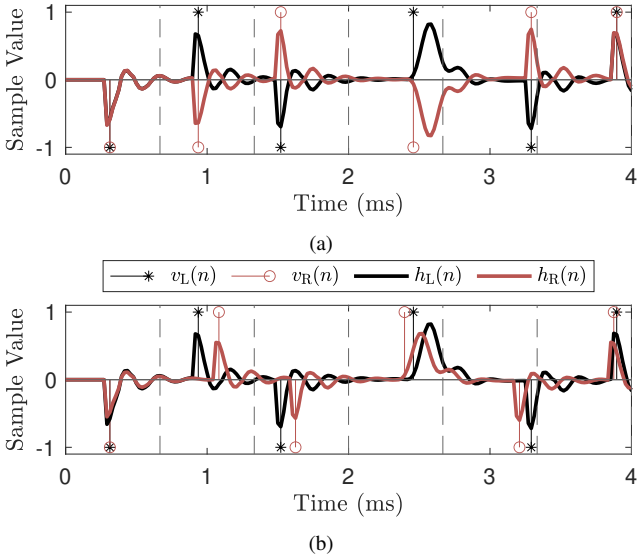


Figure 3: *(a) First four milliseconds of an incoherent two-channel EDVN IR and of (b) a BDVN IR using the jitter method. The underlying pulses are shown with stems for the right (red circle) and left (black asterisk) channels.*

3. PROPOSED BINAURAL LATE-REVERBERATOR

Building upon the previous EDVN algorithm [14], this section introduces two variants of the BDVN reverberator. The first variant directly integrates the cross-mixing approach[3], relying on synthesizing two incoherent channel sequences. The second variant achieves the desired IC by jittering the pulse locations of one channel according to a jitter distribution derived from the CPSD. A MATLAB implementation of the proposed method is available online[1].

3.1. Cross-Mixing-Based Coherence Matching

The cross-mixing requires synthesizing two incoherent sequences to provide an accurate coherence match [3]. Previous work on generating incoherent velvet-noise sequences relies on the permutation of the branches of the interleaved velvet-noise algorithm [10, 17]. In the current work, we propose the EDVN algorithm [14] (cf. Fig. 2a) to be amended to create two incoherent outputs h_1 and h_2 as shown in Fig. 2b. By transposing the original structure of Fig. 2a, the dictionary filters and pulse gains can be shared between the two channels, assuming there are no large color differences between them. The only additions are a second delay line for the second channel and the channel gains G_1 and G_2.

The two outputs h_1 and h_2 of the respective delay lines become incoherent by randomizing the signs $s(m)$ between the channels. The sign randomization is shown in Fig. 2b with the ± 1 gains. The beginning of an example incoherent two-channel EDVN IR is shown in Fig. 3a, where some of the pulse signs are flipped and some are not. Over a long sequence, the random sign flips result in incoherent signals $h_1(n)$ and $h_2(n)$. Subsequently, the two incoherent outputs are cross-mixed, following the method-

[1] https://github.com/Ion3rik/dark-velvet-noise-reverb

ology outlined in Section 2.2, utilizing the filters $U_1(z)$ and $U_2(z)$ to attain the desired IC.

3.2. Jitter-Based Coherence Matching

This section introduces the second variant of the BDVN, where the target IC is achieved by jittering the delays of the second channel based on those of the first channel. In contrast to the cross-mixing approach, the signs $s(m)$ are shared between the channels in the jitter variant. Thus, the sign flip blocks in Fig. 2b can be bypassed. By selecting appropriate jitter values, h_1 and h_2 achieve the target IC directly and thus correspond to h_L and h_R. Consequently, the cross-mixing block depicted in Fig. 2b can also be bypassed.

To find the probability mass function $p_\Delta(l)$ from which these jitter values $\Delta(m)$ are drawn, we assume — without loss of generality — that h_R is a jittered version of h_L. If the pulses, which are equivalent to the delays in Fig. 2, used to generate h_L are placed at $k_\mathrm{L}(m) = k(m)$ according to the EDVN algorithm, see Eq. (7), the pulses for h_R are placed at $k_\mathrm{R}(m) = k_\mathrm{L}(m) + \Delta(m)$, while making sure that no negative pulse locations are generated. Fig. 3b shows the beginning of an example jittered BDVN IR, where the jitter $\Delta(m)$ is visible as random time delays between the left (black) and right (red) channel pulses. The absence of the random sign flips is also evident in the figure.

The pulse sequences before applying the coloration filters are denoted here as v_L and v_R. These sequences would be obtained if all filters were set to unit impulses in Eq. (11). We utilize the fundamental relationship between the CPSD and the cross-correlation function r_LR in that the CPSD can be found as the DTFT of the cross-correlation, as in

$$S_\mathrm{LR}(\omega) = \sum_{l=-\infty}^{\infty} r_\mathrm{LR}(l) e^{-\imath\omega l}, \tag{15}$$

where the cross-correlation is

$$r_\mathrm{LR}(l) = \mathbb{E}\{v_\mathrm{L}(n) v_\mathrm{R}(n+l)\}, \tag{16}$$

in which $\mathbb{E}\{\}$ denotes the expectation operator.

For the white, zero-mean pulse sequences, the expected value is only non-zero if the shift applied to $v_\mathrm{L}(n)$ is exactly l, i.e., $\Delta(n) = l$:

$$r_\mathrm{LR}(l) = \mathbb{E}(v_\mathrm{L}(n) v_\mathrm{L}(n + \Delta(n)) \tag{17}$$

$$= v_\mathrm{L}^2(n) p(\Delta(n) = l) \tag{18}$$

$$= \sigma_\mathrm{L}^2 p(\Delta(n) = l), \tag{19}$$

assuming that the value of the sequence is independent of the jitter process. Here, σ_L^2 is the variance of the sequence, which can be set to 1.

Now $p(\Delta(n) = l) = p_\Delta(l)$ is the distribution from which the jitter values are drawn. Thus, using Eq. (15), the generated CPSD is found directly as the DTFT of the jitter distribution:

$$S_\mathrm{LR}(\omega) = \sigma_\mathrm{L}^2 \sum_{l=-\infty}^{\infty} p_\Delta(l) e^{-\imath\omega l}. \tag{20}$$

For synthesizing a sequence with a given coherence, the inverse DTFT can be used:

$$p_\Delta(l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{S}_\mathrm{LR}(\omega) e^{\imath\omega l} d\omega. \tag{21}$$

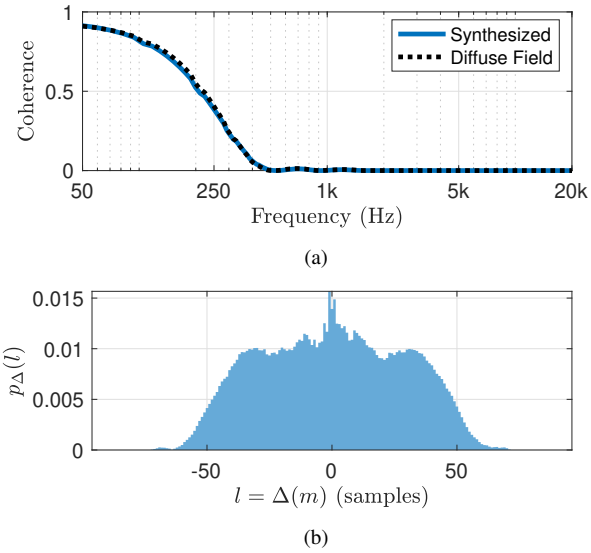


Figure 4: *(a) Jitter distribution p_Δ obtained via solving the inverse problem of Eq.* (21)*. The CPSD was analyzed from a 60-s binaural diffuse-field noise sequence, synthesized from a set of HRTF measurements. (b) The corresponding coherence estimated from the diffuse field binaural noise (dotted) and the resynthesized jittered velvet noise sequence (blue).*

The target CPSD $\hat{S}_\mathrm{LR}$ used here should not simply be the numerator of Eq. (1). Coloration should not be modeled using the jitter distribution, as it is modeled separately by the dictionary filters. Therefore, whitening should be applied to the estimated CPSD. We used

$$\hat{S}_\mathrm{LR}(\omega) = \frac{\sum_{\nu=\nu_0}^{\infty} H_\mathrm{L}(\nu,\omega) H_\mathrm{R}^*(\nu,\omega)}{\sqrt{\sum_{\nu=\nu_0}^{\infty} |H_\mathrm{L}(\nu,\omega)|^2 \sum_{\nu=\nu_0}^{\infty} |H_\mathrm{R}(\nu,\omega)|^2}}, \tag{22}$$

to design the jitter distribution, the square of which is the coherence.

Once the distribution is obtained using Eq. (21), the jitter samples for each pulse $\Delta(m)$ need to be generated. Random samples from a unidimensional distribution can be drawn with the inversion method, drawing samples from a uniform distribution and mapping those values via the inverse cumulative density function (CDF) of the random variables, i.e.,

$$\Delta(m) = F_\Delta^{-1}(r_3(m)), \tag{23}$$

where $r_3(m)$ is a uniform random variable between 0 and 1, and F_Δ^{-1} is the inverse CDF of p_Δ, obtained numerically.

As an example, Fig. 4a shows the binaural diffuse field coherence of the KU100 dummy head [18], estimated from binaural noise generated via filtering incoherent, white Gaussian noise with a set of measured HRTFs in a quasi-uniform grid. It also shows the synthesized coherence, which exhibits a close match to the target diffuse field coherence. Fig. 4b shows the jitter distribution that was determined from the diffuse field coherence, with Eq. (21), and used for synthesis.

Table 1: *Target BRIR parameters as averages between 500 Hz and 1 kHz, following the standard ISO 3382. The Direct-to-reverberant ratio (DRR) is shown for left and right ear responses. "Arni" refers to the variable acoustics room at the Aalto Acoustics Lab.*

Target BRIR	Description	T20 (s)	DRR (dB)
Room 1	Arni Medium	0.49	-7.6 / -2.8
Room 2	Arni Reverberant	0.98	-10.0 / -6.5
Room 3	Pori Promenadi Hall	2.24	-19.0 / -18.0

4. EVALUATION

This section presents both objective and subjective assessments of the proposed BDVN algorithm. Specifically, the proposed method was used to model the binaural coherence of three distinct measured BRIRs, alongside room-independent binaural diffuse field coherence derived from a collection of HRTF measurements. Furthermore, the capabilities of the BDVN algorithm in synthesizing altered coherence patterns beyond those inherent in natural binaural listening scenarios, are explored. This includes discussing the algorithm's efficacy in implementing parametric width control for coherence and its ability to generate diverse time-dependent coherence modifications. Sound examples are provided online for the curious reader to experience the quality of the BDVN algorithm themselves [2].

4.1. Target Coherence Profiles

Three different BRIRs with varying reverberation times were selected to be modeled with the two proposed method variants. Table 1 shows the reverberation time and direct-to-reverberant ratio (DRR) estimated from the target BRIRs. The selected rooms from Room 1 to 3 are ordered from the driest (Room 1) to the most reverberant (Room 3). "Room 1" and "room 2" correspond to measurements that were both made in the variable acoustics room "Arni", at the Aalto Acoustics Lab, in which the room was configured to two different settings. In both cases, the source was 3.3 m in front and 1.3 m to the right of the binaural receiver (a KEMAR head and torso simulator). The fact that the source was not central can be seen in the increased DRR at the right ear.

All the coherence estimates and the BDVN modeling were implemented for the late part (after 50 ms) of each BRIR. In addition to the target coherence estimated from the varied set of BRIRs, the binaural diffuse field of the KU100 shown above is included as one of the target coherences. The motivation is to investigate whether room-independent binaural coherence provides a perceptually accurate approximation for each tested room. It is expected that different head sizes lead to slightly different IC, but studying these differences is beyond the scope of this paper.

The coherence of the selected BRIRs and the diffuse field binaural coherence is shown in Fig. 5. It can be seen that the coherence of the measured BRIRs is always larger or equal to that of the binaural diffuse field coherence, across all frequencies. This is expected as rooms are not expected to generate perfect diffuse fields (for example due to anisotropy [19, 20]). In general, the coherence of each BRIR still follows the binaural diffuse field coherence fairly closely.

[2] http://research.spa.aalto.fi/publications/papers/dafx24-bdvn/

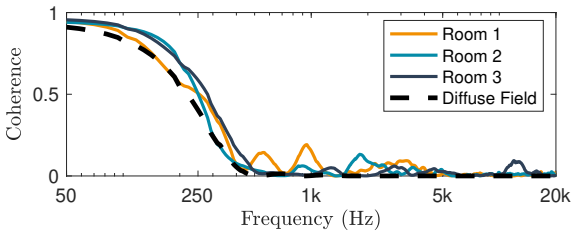


Figure 5: *Binaural diffuse field coherence and coherence of the three BRIRs beyond 50 ms.*

4.2. Objective Evaluation

The spectrograms of the three target BRIRs (top) and the jitter-based BDVN model instances (bottom) tuned to the target BRIRs are shown in Fig. 6. Each model instance includes the original direct and early part from the target BRIR and the modeled late reverberation part. The dictionary filters and the pre-and post-filters for the three cases were estimated from the left channel of each target BRIR, as no large spectral differences were identified between the left and right channels in Fig. 6. More details are provided on fitting the model to a target response in [14]. Each BDVN model instance utilized a pulse density of $\rho = 1500$, $Q = 10$ fifth-order allpole dictionary filters, a single 10th-order post-filter, and a second-order DC-blocker. The computational costs of the method are analyzed in more detail in [13].

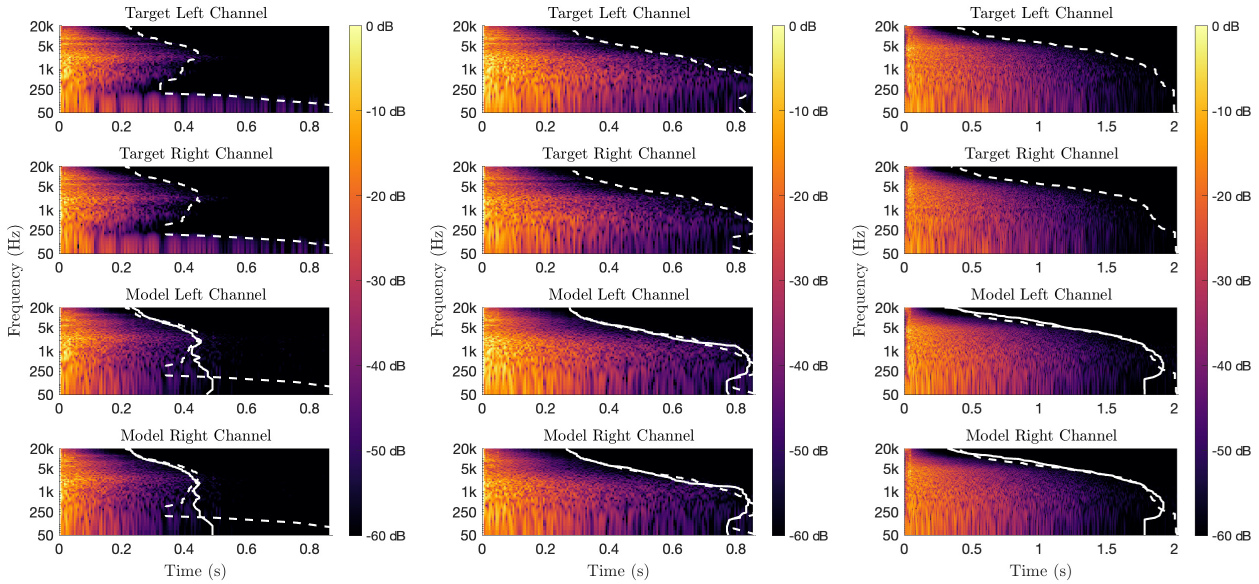
The reverberation time (RT) estimates of the target BRIRs (dashed) and the model BRIRs (solid) are overlayed on the spectrograms. The median RT error over all frequencies is 3.2%, 1.7%, and 7.1%, for Rooms 1, 2, and 3, respectively. A large error between the target and model RT can be observed for Room 1 at the low frequencies (cf. Fig. 6a). However, the error arises mostly from the DC-blocker of the model which is by design removing low-frequency noise present in the target BRIR [14].

Three different BDVN models with different coherence-matching approaches were generated for each of the three target BRIRs. The coherence estimates are shown in Fig. 5. The incoherent version refers to using h_1 and h_2 of Fig. 2b directly when the signs of the channels are randomized. The cross-mixed version (BDVN mix) is then derived from h_1 and h_2 via applying the cross-mixing stage and taking the output from h_L and h_R. Finally, two versions utilizing the jitter approach were generated; one fitted to each measured coherence separately (BDVN jit.), and one fitted to the generic binaural diffuse field coherence (BDVN jit. diff.).

The driest room, Room 1, shows some increase in coherence due to the limited length of the underlying sequences for the incoherent version Fig. 7a (green). Room 2 in Fig. 7b shows the largest difference between the binaural diffuse field match (BDVN jit. diff.) and the measurement matches (BDVN jit. and BDVN mix). For Room 3 (Fig. 7c) all the versions show a slightly lower coherence at lower frequencies compared to the measured coherence (Ref).

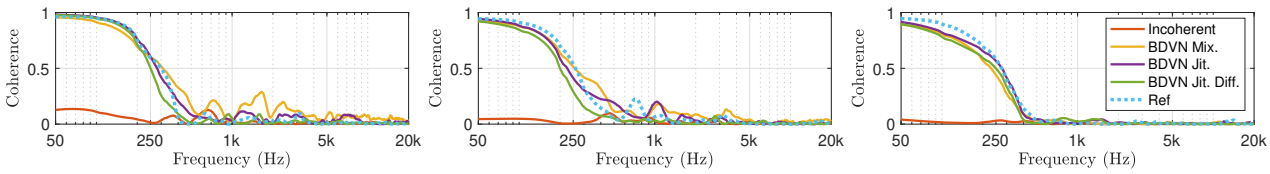
4.3. Perceptual Test

A Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) like test was conducted to evaluate the similarity between modeled and measured BRIRs. In total, 16 participants took the test (mean age 27.6 years, standard deviation 3.9 years). All of



(a) (b) (c)

Figure 6: Spectrograms of the three measured BRIRs (top) and the corresponding BDVN jit. model instances (bottom) of (a) Room 1, (b) Room 2, and (c) Room 3. The reverberation time estimates of the target BRIRs (dotted) and BDVN model instances (solid) are overlayed on the spectrograms.



(a) (b) (c)

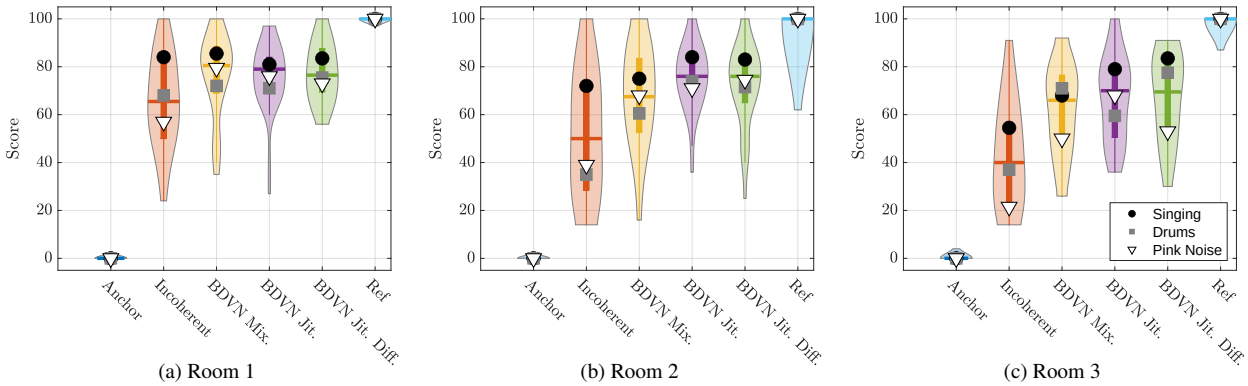
Figure 7: Coherence of BDVN model instances tuned to (a) Room 1, (b) Room 2, and (c) Room 3.

the participants reported having normal hearing and 15 of them had previous experience with formal listening tests. The listening test was implemented with the webMUSHRA platform [22], and conducted in the sound-proofed listening booths at the Aalto Acoustics Lab using Sennheiser HD-650 headphones.

The task in the test was to assess the similarity of renderings using artificial reverberation to a reference measured BRIR. The five conditions discussed above, and their coherence shown in Fig. 5, were used. In addition, monaural reverberation, which was the right channel of BDVN jit., served as an anchor, yielding six conditions in total. The test signals included male singing, a drum loop, and pink noise. With these samples, we aimed to answer the following three questions: 1) Does applying binaural coherence yield an improvement over using incoherent channels? 2) Is there a difference between the cross-mixing- and the jitter-based BDVN? 3) Is fitting the coherence to each room necessary, or can the diffuse field coherence be used for all rooms?

Four participants were excluded from the analysis since they gave a rating lower than 95 for the reference in more than 15% of the trials. The results are shown in Fig. 8, where different signals are marked by different symbols. In general, the highest scores were obtained using singing, especially in Room 1.

For statistical analysis, non-parametric tests were used, since all samples could not be considered stemming from a normal distribution. Disregarding the anchor and the reference, Friedman tests indicated significant differences between conditions in all three rooms, with Chi-squared values of 17.77, 26.55, and 44.55 for Rooms 1, 2, and 3, respectively. All of these lead to $p < 10^{-3}$. Pairwise comparisons using Wilcoxon signed rank tests with the Bonferroni-Holm correction revealed significant differences between incoherent noise and the three BDVN conditions. For Room 1, the differences were the smallest. Here, the differences in median between incoherent generation and the BDVN condition with the lowest performance (BDVN - Jit. Dif.) was 11 points, $Z = -3.5$, $p < 0.001$. In the other two rooms, the differences between incoherent EDVN and all BDVN variants were larger; for Room 2, the difference between incoherent generation and BDVN Mix (the lowest performing BDVN alternative in this case) was 17.5 points, $Z = -3.09$, $p = 0.002$; for Room 3 the difference between incoherent signals and BDVN Mix was 26 points, $Z = -3.83$, $p < 0.001$. Thus, using BDVN consistently improves the match compared to using incoherent signals for both



(a) Room 1 (b) Room 2 (c) Room 3

Figure 8: *Perceptual test results for the three rooms and three signals* `violinplot` *function [21]. The boxplot shape is included as a black line in the center of the violin. The large black dot, gray square, and white triangle indicate median results for singing, drums, and pink noise, respectively. The horizontal lines indicate the overall medians and the bottom and top edges of the boxes indicate the 25th and 75th percentiles, respectively. The violin outline shows the kernel density estimation.*

ears.

Comparing the BDVN variants to each other, some trends, but no significant differences were found. Interestingly, for Room 1, all methods performed equally well, with median scores of 80.5 for BDVN Mix, 79.0 for BDVN Jit., and 76.5 for BDVN Jit. Diff. For Rooms 2 and 3, the BDVN with cross-mixing performed slightly worse than the jitter approach. For Rooms 2 and 3, BDVN with the jitter approach fit to the room and using diffuse field coherence yielded almost identical results (medians both 76.0 for Room 2 and 70.0 vs 69.5 for Room 3). Thus, matching the diffuse field coherence yields favorable results for the given rooms, particularly in the more reverberant rooms 2 and 3.

4.4. Beyond Natural Coherence

Besides replicating natural binaural coherence in physical rooms, the jitter-based BDVN variant facilitates the generation of artistic two-channel reverberation effects. These include a parametric width control and various time-dependent effects. Fig. 9a shows a parametric Hann-window distribution, with various maximum jitters. Changing the maximum jitter of the distribution alters the IC as shown in Fig. 9b, which in turn alters the perceived width of the reverberation. Increasing the maximum jitter reduces coherence, notably shifting the IC cutoff towards lower frequencies. Furthermore, by adjusting the selected parametric distribution, tuning the maximum jitter can effectively match the binaural diffuse field coherence (dashed line).

The maximum jitter can be time-dependent without extra computational load, enabling reverberation with time-varying coherence and thus perceived source width. Fig. 10a shows a coherence profile that becomes more and more coherent towards the end of the BDVN response. The effect is implemented by decreasing the maximum jitter (i.e., the width of the jitter distribution of Fig. 9a) in time. This creates an unnatural collapsing sensation where the perceived width of the late reverberation starts wide and then collapses in the middle. Another possible time-dependent coherence effect is presented in Fig. 10b, where the maximum jitter is modulated with a low frequency (2 Hz) sine wave. The resulting perceptual effect is a modulated panning-like sensation in the reverberant tail, as the two channels move between incoherent and coherent.

5. CONCLUSIONS

BDVN is proposed in this paper as a two-channel extension of the previous EDVN algorithm, capable of synthesizing binaural coherence between two output channels. The desired coherence can be generated via cross-mixing and filtering two incoherent channels or with the jitter-based approach. For the latter, the jitter distribution derivation based on a target coherence profile is presented, and a numerical simulation is provided to confirm the method's effectiveness. The jitter and cross-mixing-based BDVN instances were parametrized to model three different BRIRs and compared objectively and subjectively. The objective evaluation shows that the coherence of each target BRIR can be matched accurately with both alternatives. Furthermore, it is shown that each of the measured BRIRs' coherence is similar to a generic, diffuse field binaural coherence.

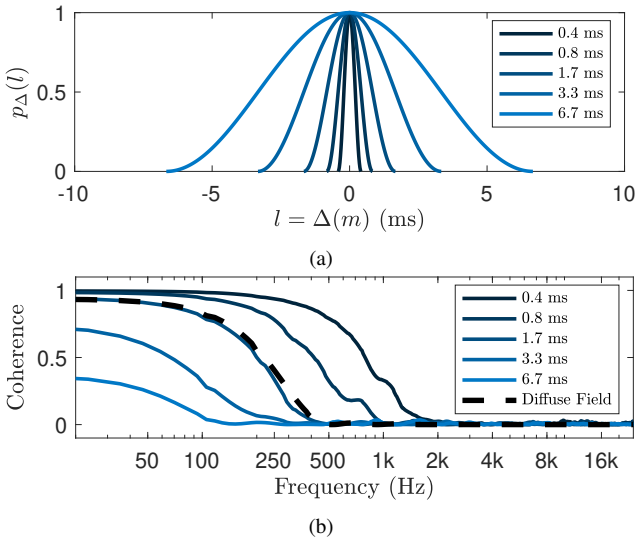
The perceptual test revealed a significant difference between incoherent rendering and the BDVN methods. The cross-mixing and jitter-based methods showed no significant differences between each other. Moreover, similar scores were obtained when matching the room-specific coherence or the generic binaural diffuse field coherence, suggesting that a room-independent static binaural rendering might be a perceptually accurate model for synthesizing binaural late reverberation of any room.

Finally, the benefits of the jitter-based method over cross-mixing were highlighted by introducing a parametric width control for the jitter-based approach. Additionally, it was shown that a time-dependent jitter distribution can be designed with no added computational load, generating various exciting artistic reverberation effects beyond the natural reverberation. These effects presented here include widening or narrowing the width in time and applying arbitrary IC modulation.

6. REFERENCES

[1] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Proc. AES 98th Conv.*, Feb. 1995.

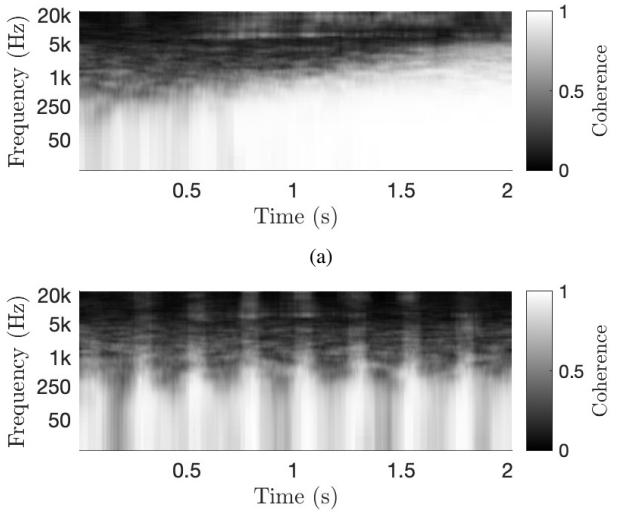
[2] F. Menzer and C. Faller, "Investigations on modeling BRIR



(a)

(b)

Figure 9: (a) Hann-window jitter distribution with various maximum jitters and (b) the corresponding coherence. The dotted line shows the diffuse field binaural coherence.



(a)

(b)

Figure 10: (a) Increasing coherence perceived as the reverb image collapsing towards the middle and (b) sine-modulated coherence generating a subtle panning-like movement in the reverberation.

tails with filtered and coherence-matched noise," *J. Audio Eng. Soc.*, 2009.

[3] F. Menzer and C. Faller, "Binaural reverberation using a modified Jot reverberator with frequency-dependent interaural coherence matching," in *Proc. 111th AES Conv.*, Munich, Germany, Jan. 2009, paper 7765.

[4] F. Menzer, "Binaural reverberation using two parallel feedback delay networks," in *Proc. AES 40th Int. Conf. Spatial Audio: Sense the Sound of Space*, Oct 2010.

[5] C. Kirsch, J. Poppitz, T. Wendt, S. van de Par, and S. D. Ewert, "Computationally efficient spatial rendering of late reverberation in virtual acoustic environments," in *Proc. Immersive and 3D Audio (I3DA)*, Sep. 2021, pp. 1–8.

[6] N. Agus, H. Anderson, J.-M. Chen, S. Lui, and D. Herremans, "Minimally simple binaural room modeling using a single feedback delay network," *J. Audio Eng. Soc.*, 2018.

[7] C. Kirsch, J. Poppitz, and T. Wendt, "Spatial resolution of late reverberation in virtual acoustic environments," *Trends in Hearing*, vol. 25, 2021.

[8] M. Karjalainen and H. Järveläinen, "Reverberation modeling using velvet noise," in *Proc. 30th AES Int. Conf. Intelligent Audio*, Saariselkä, Finland, Mar. 2007.

[9] K. Lee, J.S. Abel, V Välimäki, T Stilson, and D. P. Berners, "The switched convolution reverberator," *J. Audio Eng. Soc.*, vol. 60, no. 4, pp. 227–236, Apr. 2012.

[10] V. Välimäki and K. Prawda, "Late-reverberation synthesis using interleaved velvet-noise sequences," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1149–1160, Feb. 2021.

[11] B. Holm-Rasmussen, H.-M. Lehtonen, and V. Välimäki, "A new reverberator based on variable sparsity convolution," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Maynooth, Ireland, Sep. 2013, pp. 344–350.

[12] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H-M. Lehtonen, "Late reverberation synthesis using filtered velvet noise," *Appl. Sci.*, vol. 7, no. 5, May 2017.

[13] J. Fagerström, N. Meyer-Kahlen, S. J. Schlecht, and V. Välimäki, "Dark velvet noise," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Vienna, Austria, Sep. 2022, pp. 192–199.

[14] J. Fagerström, S. J. Schlecht, and V. Välimäki, "Non-exponential reverberation modeling using dark velvet noise," *J. Audio Eng. Soc.*, vol. 72, no. 6, pp. 370–382, Jun. 2024.

[15] P. Goetz, K. Kowalczyk, A. Silzle, and E. Habets, "Mixing time prediction using spherical microphone arrays," *J. Acoust. Soc. Am.*, vol. 137, pp. 206–212, Feb. 2015.

[16] N. Meyer-Kahlen, S. J. Schlecht, and V. Välimäki, "Colours of velvet noise," *Electron. Lett.*, vol. 58, no. 12, pp. 495–497, Jun. 2022, https://doi.org/10.1049/ell2.12501.

[17] K. Prawda, S. J. Schlecht, and V. Välimäki, "Multichannel interleaved velvet noise," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Vienna, Austria, Sep. 2022, pp. 208–215.

[18] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU100," 2013.

[19] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual analysis of directional late reverberation," *J. Acoust. Soc. Am.*, vol. 149, no. 5, pp. 3189–3199, May 2021.

[20] M. Berzborn and M. Vorländer, "Directional sound field decay analysis in performance spaces," *Building Acoustics*, vol. 28, no. 3, pp. 249–263, Sep. 2021.

[21] B. Bechtold, "Violin Plots for Matlab, Github Project," Available at https://github.com/bastibe/Violinplot-Matlab, accessed Mar 27, 2024.

[22] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "WebMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, pp. 1–8, Feb. 2018.